Robot-Mediated Assistance: Opportunities and Challenges in Computer Vision and Human-Robot Interaction

MD ALIMOOR REZA, Department of Mathematics and Computer Science, Drake University, USA
SYED MASUM BILLAH, College of Information Science and Technology, Pennsylvania State University, USA

Robot-mediated assistance is poised for wide acceptance in sectors such as households, healthcare, distance learning, and assisted living in the near future. In this work, we analyze a curated collection of 29 existing robots. This analysis leads us to extend the current human-robot interaction (HRI) taxonomy by incorporating three additional lenses: robot appearance (including human-like, animal-like, and machine-like), collaborators (involving local users, bystanders, tele-users, tele-operators, other robots, and AI developers), and abilities (such as locomotion, perception, grasping, manipulation, and communication with humans or robots). We examine how these factors influence the quality, utility, and nature of collaboration in robot-mediated interaction. Further, we scrutinize the challenges and opportunities in robotics, computer vision, and HRI, aiming to stimulate and inform future research in these areas.

ACM Reference Format:
Md Alimoor Reza and Syed Masum Billah. 2023. Robot-Mediated Assistance: Opportunities and Challenges in Computer Vision and Human-Robot Interaction. 1, 1 (August 2023), 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Tags: Robot-mediated assistance, service robots, assistive robots, tele-presence robots, tele-assistance, remote assistance; artificial intelligence, computer vision; human-robot interaction, human-computer interaction;

1 INTRODUCTION

Robot-mediated assistance, wherein robots increasingly assume pivotal roles as collaborative partners with humans, stands as a technological breakthrough on the brink of reshaping society. Robots, proving indispensable across diverse contexts [60, 73, 101], range from aiding in assisted care to handling routine tasks. For instance, in the *general home service* sector, robots undertake tasks like clearing the dinner table, doing laundry, or engaging in crafting activities. In *assisted living service*, they assist individuals by grasping canes, navigating environments, and performing various support tasks. Additionally, in the realm of *tele-medicine service*, robots can interpret medicine labels and retrieve specified medications. Depending on the nature of their tasks, these robots operate from fixed locations—like workstations or bedside setups—or navigate within a defined environment. Their autonomy ranges from independent navigation and semi-autonomous maneuvering to tele-operation under human supervision.

Two primary categories typify robots based on their roles and modes of human interaction: *social* and *service* robots. Social robots are engineered to engage in socially meaningful interactions with humans, while service robots concentrate on offering functional assistance to simplify and expedite the execution of physical tasks. Service robots are rapidly evolving to serve citizens more effectively in the future [144]. Cases in point include the tele-operation of a robotic arm

Authors' addresses: Md Alimoor Reza, Department of Mathematics and Computer Science, Drake University, USA, md.reza@drake.edu; Syed Masum Billah, College of Information Science and Technology, Pennsylvania State University, USA, skb5969@psu.edu.

via virtual reality (VR) to set a dining table from halfway across the globe [161], and the automatic synthesis of virtual wheelchair training environments [89]. Such advancements can not only blur geographical boundaries but also extend the reach of service robots beyond physical constraints, enhancing human skills and quality of life [28]. Concurrently, social robots are beginning to find applications in sectors such as mental healthcare and education [120, 121].

In traditional tele-assistance settings, a human provider remotely aids another human receiver through audio and video communication. However, introducing a robot into this framework ushers in a new paradigm of robot-mediated tele-assistance. In this arrangement, a remote human provider can instruct a robot to perform various tasks such as picking out the correct medicine bottle from a cabinet for an older adult, locating dropped objects like keys or wallets for a visually impaired user, or serving food to patrons at a restaurant table.

Four critical dimensions characterize robot-mediated interactions. Firstly, while humans and robots form a team and work jointly towards a shared goal, the actions available to each team member may vary [30]. For example, robots are well-suited to manage dangerous or unpleasant tasks under remote human operations. Secondly, the responsibilities of each team member can shift as a task progresses, fostering dynamic and adaptive human-robot collaboration [27]. Thirdly, such collaborations may span considerable durations, with repeated interactions fostering shared knowledge and conventions such as role specialization based on skill recognition [30, 132]. Lastly, collaborations can also be brief, with the human partner frequently rotating.

In the exploration of robot-mediated interaction, we can scrutinize the roles of humans and robots from multiple perspectives. One foundational classification, as proposed by Yanco et al., provides broad categories of interaction [157].

- *Autonomy and Intervention Levels*: These measure the percentage of time the robot performs tasks independently and the percentage of time it requires human intervention, respectively. The two sums to 100%, denoting the division of task completion between robotic autonomy and human involvement.
- *Space-Time Taxonomy*: This approach classifies human-robot interaction based on its synchronicity (simultaneous or time-separated) and location (same or different places).
- *Composition of Robot Teams*: Multiple robots can form either homogeneous teams (e.g., multiple Roombas cleaning a house) or heterogeneous teams (e.g., an Astro playing with a pet before a Roomba cleans up).
- *Human-to-Robot Ratio*: This simply counts the number of humans and robots involved (e.g., 1:2 represents one human operating two robots).
- *Shared Interaction*: The level of shared interaction can vary from one human interacting with a single robot or a team of robots to multiple humans interacting with a single robot.
- *Decision Support*: Factors such as available sensor information, sensor fusion type, and pre-processing can facilitate human-robot interaction.
- *Task Criticality*: This assesses the importance of accurately completing a task and the potential negative effects of failure, categorized as high, medium, or low.

While this taxonomy offers a comprehensive framework, it leaves out certain aspects of robot characteristics and functionalities, such as appearance, capabilities, and collaborative partners. For instance, a robot's appearance can significantly influence its reception and effectiveness [23]. Similarly, factors contributing to a robot's autonomy, such as locomotion, perception, communication, and manipulation abilities, are not distinctly highlighted. To address this gap, we examine a sample of 29 existing robots (Table 1). Our selection of these robots is informed by prevalent trends in AI and Human-Robot Interaction (HRI) literature, industry movements, and relevant newsletters (e.g., IEEE Spectrum Robotics). Based on our analysis, we propose the following three additional lenses:

- *Robots' Appearance*. We consider the varying appearances of robots, including human-like, machine-like, animal-like, and combinations of these appearances (Section 2).
- *Robots' Collaborators*. We look at the diverse roles humans can assume in collaboration with robots, such as users, tele-operators, bystanders, and robot developers (Section 3).
- *Robots' Abilities*. We examine the broad range of robotic capabilities, including locomotion, perception, grasping and manipulation, and communication with humans and other robots (Section 4).

This chapter is organized as follows: Section 2 describes the crucial role that the physical appearance of a robot plays in shaping human-robot interactions. The challenges presented in this section revolve around designing a robot's appearance to accurately reflect its behavioral capacities, operational domain, and intended functionality. The importance of striking a balance between the function and design of a robot's appearance is emphasized. The section also reveals a promising path towards creating robots with hybrid and shape-changing appearances. Such adaptive robots, capable of altering their form based on situational needs, could cater to specific user requirements, marking an exciting direction for future research in robotics and human-robot interaction.

Section 3 details several distinct roles within the robot-mediated assistance framework: the robot, user, bystanders, tele-operator/tele-user, AI developer, environments, and providers. The majority of robots are designed for single-user tasks, limiting their potential. Introducing a tele-operation interface could enhance their utility, but it poses challenges in observing social norms and privacy. Coordinating multiple robots or users presents difficulties due to limited interaction possibilities. Robots designed for interaction with other robots lack a uniform communication protocol. Allowing end-users to reprogram robots is crucial for improvement. Integrating bystanders' input can enhance a robot's task performance, offering a significant opportunity.

Section 4 explores the opportunities and challenges associated with the robots' abilities, such as locomotion, perception, grasping and manipulation, and communication with humans and other robots. Robots' locomotion (Section 4.1) capability addresses their movement from one location to another. While some robots are anchored to a specific platform, many assistive applications require autonomous or semi-autonomous movement. Robots can employ various locomotion mechanisms, including crawling, sliding, or walking. They typically use wheeled, chain/track, or legged systems, each with hardware design and software challenges to address. Robotic perception (Section 4.2.1) relies on various computer vision tasks, such as object recognition, object tracking, and segmentation. Opportunities involve the use of active learning, human-in-the-loop techniques, and other robust approaches. Challenges include handling lighting variations, occlusions, and dynamic environments. Further, robotic grasping and manipulation (Section 4.3) in unstructured environments pose challenges in selecting suitable configurations and performing varied tasks. Opportunities include developing universal object-picking robots, improving algorithms and hardware, and facilitating adaptive tool use for enhanced real-world assistance. Finally, the communication between humans and robots (Section 4.4) poses key challenges in Human-Robot Interaction (HRI). These include fostering long-term engagement, managing dependence on robots, and ensuring cultural sensitivity. Creating enduring human-robot relationships, and designing culturally sensitive robots are critical. Deciphering human non-verbal cues and gestures is challenging but crucial for nuanced interactions. Security, privacy, and trust are also significant concerns in HRI. Evaluating HRI is complex due to the scarcity of open robot platforms, the lack of real-world datasets, and limited access to evaluation settings. Innovative interaction modalities, such as Virtual Reality/Augmented Reality (VR/AR) technologies and Brain-Control Interfaces (BCIs), have the potential to enable shared and embodied interactions in HRI.

2 ROBOTS' APPEARANCE

A robot's appearance, encompassing its form, structure, and visual characteristics, greatly influences human-robot interactions. Robots designated for social roles often showcase anthropomorphic features, which users tend to prefer [83, 93, 154]. However, overly intricate designs may inflate user expectations, resulting in diminished satisfaction when these expectations remain unfulfilled [15, 39, 113]. We categorize robot appearances into three main groups:

- *Human-like:* Robots bearing human-like features such as a face, arms, legs, or even a complete exoskeleton. These robots operate in diverse environments and possess the ability to navigate various terrains.
- *Animal-like:* Robots resembling animals, equipped with multiple legs or arms. Like their human-like counterparts, these robots can navigate an array of terrains, including flat surfaces, stairs, and uneven ground.
- *Machine-like:* Robots that emulate man-made machinery often incorporate wheels or tracks for movement, thus allowing for efficient traversal on flat surfaces.

2.1 Challenges and Opportunities in Robot Appearance

Mori's uncanny valley theory [106] offers a valuable guideline regarding robot appearance. The theory proposes a critical point of human-likeness in robots, where positive emotional responses suddenly give way to aversion. However, if the human-likeness continues to escalate beyond this point, positive responses resume, approaching levels of empathy typically associated with human-human interactions. Given that the current state of robot autonomy and appearance falls short of human-like, reducing a robot's human-like appearance and behavior might prove beneficial in managing user expectations [56]. Therefore, it becomes a challenge to align a robot's appearance with its social and behavioral capacities, operational domain, and intended function.

Earlier generations of manufacturing robots adopted machine-like forms and were employed for tasks such as moving items, assembling components, and inspecting products. Today's advancements have led to a hybrid design approach, combining human and machine-like features, as seen in models like Baxter (refer to Table 1). In the service sector, robots designed to assist humans often incorporate human-like features. Examples include Pepper and NAO, which serve in classrooms and hospitality roles; Relay for transporting objects in restaurants; Moley for kitchen assistance; and Astro for home monitoring (see Table 1). The human-like faces of these robots create high expectations for their dialogue capabilities, leading to reduced interaction when these expectations remain unmet. Telepresence robots, designed to represent the user remotely, also commonly feature a human-like appearance. Although anthropomorphism is not necessary for effective teleoperation [34], the degree of anthropomorphism can influence user perceptions of presence in a virtual environment [113]. In healthcare, animal-inspired robots such as PARO (shown in Table 1) have proven effective due to their lower expectation levels. These instances underscore the importance of achieving a balance between the function and design of a robot's appearance.

Most robots exhibit either human- and machine-like attributes (e.g., 13 out of 29 robots in Table 1), followed by those purely machine-like (8), human-like (4), or animal-like (4). However, an intriguing opportunity resides in exploring hybrid designs that combine human- and animal-like features (e.g., a snake with a human face), or in investigating the potential of shape-changing robots that adapt their form based on context (e.g., adopting human-, animal-, or machine-like appearances as required).

Such adaptive, shape-changing robots could comprise reprogrammable, Lego-like modules, and assume task-specific configurations, such as robotic limbs to assist individuals with disabilities. Additionally, the form of the robot could be tailored to specific personal needs, enabling it to perform various tasks like opening and closing doors and windows,

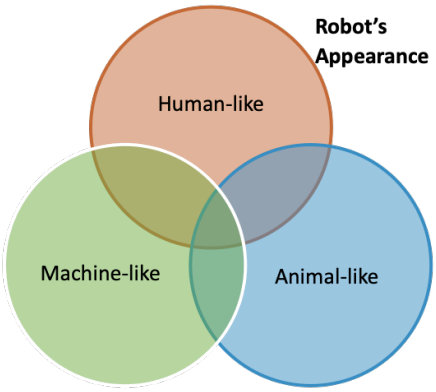

Fig. 1. Robots' appearance lens.

Fig. 2. NAO: a human-like robot (photo from [147]).

Fig. 3. AVA: a human- and machine-like robot (photo from [163]).

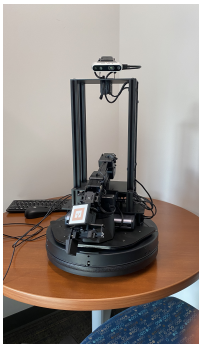

Fig. 4. LoCoBot: a machine-like robot.

Fig. 5. Paro: an animal-like robot. It looks like a pet seal (photo from [13]).

Fig. 6. Baxter: an industrial human- and machine-like robot (photo from [62]).

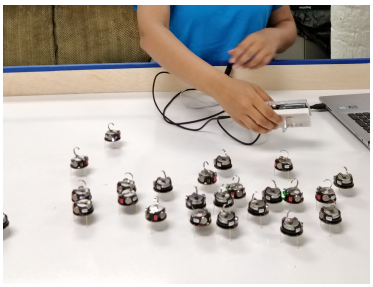

Fig. 7. Kilobots: a group of animal (insect)-like robots (photo from [119]).

forming a staircase-like shape to aid users with motor impairments, or locating dropped objects for users with vision impairments. The integration of swarm robotics, akin to Kilobot in Table 1, could facilitate this on-demand adaptability of robot appearance and functionality. This advancement in robot design presents a compelling direction for future research in robotics and human-robot interaction.

3 ROBOT'S COLLABORATORS

The robot-mediated assistance paradigm encompasses several distinct roles for individuals who can interact with the robot, as shown in Figure 8. These roles are inspired by the design challenges and guidelines for social interacting using mobile telepresence robots [145].

- *Robot*: This term straightforwardly refers to the robot itself.
- *User*: The user, who is co-located with the robot, actively interacts with and benefits from the robot's services.
- *Bystanders*: These are individuals physically present in the same environment as the user and the robot, who may interact with the robot directly or indirectly.

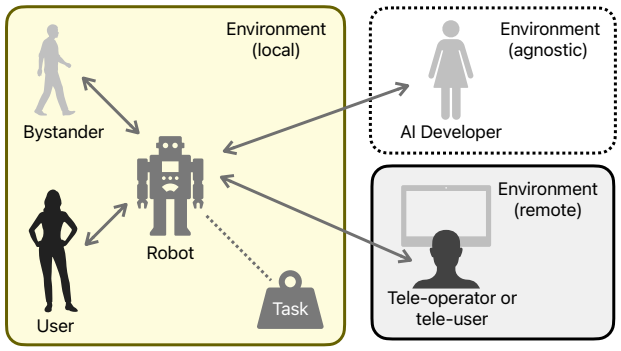

Fig. 8. Robots' collaborators' lens.

- *Tele-operator or Tele-user*: A tele-operator remotely controls the robot to serve the user through a specific human-robot interface. In cases where the tele-operator uses the robot for personal needs, such as remote conference attendance, we refer to this role as the tele-user.
- *AI Developers*: These individuals can build, install software and firmware, and test and debug the performance of the AI modules in the robot, either frequently or infrequently, locally or remotely.
- *Providers*: For the sake of simplicity, we employ this term to denote non-user human entities, such as bystanders, tele-operators, and AI developers.

We distinguish between two types of environments: the local environment where the robot, user, and bystanders are co-located; and the remote environment from where the tele-operator or tele-user controls the robot.

3.1 Challenges and Opportunities in Robots' Collaborators

A majority of robots are designed to serve a single user, either local or remote, for a specific task, as illustrated in Table 1. This design philosophy can limit the robot's potential because introducing a tele-operation interface could allow remote human operators to assist local users through the robot, particularly in complex, unstructured tasks. Implementing such an interface would thus enhance the robot's utility.

However, incorporating a tele-operation interface presents its own set of difficulties. For instance, observing social norms regarding personal space (proxemics) is critical for creating comfortable human-robot interactions. Robots should also follow social etiquette, such as preserving private communication and avoiding unauthorized audio or video recording. Moreover, robots could be designed to disclose user status and identification to promote transparency in interactions. Anthropomorphic robots, in particular, could maintain eye contact, facilitate joint attention, and adjust their height to sit or stand in alignment with the context, thus emulating human behavior [145]. We provide a more detailed discussion on this challenge in Section 4.4.

Another difficulty lies in coordinating various collaborators. Consider the scenarios where a single user owns multiple robots (e.g., a Roomba and an Astro), or a single robot serves multiple users (as when a service robot delivers a sandwich to one person and a salad to another in a care facility). At present, such coordination relies primarily on audio and video connections, with insufficient support for natural language, gestures, facial expressions, speech, and various body

and posture cues [20, 67, 145]. This limitation results in unequal interaction possibilities for different collaborators, reducing the overall utility of robot-mediated assistance.

Robots intended for interaction with other robots face an added challenge in achieving uniform communication protocol, given that many robots do not provide a Software Development Kit (SDK). Current inter-robot communication is ad hoc, and for those robots with SDK, only AI developers can update firmware. To improve this situation, it is crucial to enable end users to reprogram their robots, an action currently restricted to programming.

Lastly, it is noteworthy that most robots regard bystanders merely as obstacles to be avoided for safety. However, bystanders can offer valuable input to robots to enhance their task performance. Therefore, integrating bystanders into a robot's decision-making process presents another significant opportunity.

4 ROBOTS' ABILITIES

The abilities of robots are most accurately characterized by Moravec's paradox [105], which posits that tasks humans find simple, such as sensorimotor activities and perception, tend to be challenging for robots. On the other hand, tasks that are often demanding for humans, like computations, are more readily performed by robots. This paradox is deeply rooted in human evolution [105]. Through the process of natural selection, the sensory and motor regions of the human brain have undergone significant evolution, enabling humans to effectively engage with the natural world and survive within it. Therefore, to reverse-engineer these biologically refined and seemingly effortless human skills into robots is expected to be challenging [102, 105].

In the current landscape of robotics, human-like skills are grouped into three fundamental categories: **sensing**: the processing of sensor-derived information; **planning**: the computation of suitable directives; and **actuation**: the generation of commands for mechanical actions. Depending on the tasks, robots may also require **communication** skills for interaction with humans or other robots, and might need to exhibit **locomotive** abilities. Arranging these skills differently yields various robot abilities. In this section, we identify four such abilities, each discussed further below, along with their respective challenges and opportunities.

- *Locomotion*: This pertains to the robot's navigational ability, specifically moving from point A to point B. Key sensors and hardware for locomotion include wheels, legs, and chains.
- *Sensing or Perception*: This domain involves interpreting visual scenes, decoding natural language and speech, and formulating plans. Necessary sensors and hardware for perception encompass RGB cameras, depth sensors (like LiDAR or IR sensors), bump sensors, tactile sensors, temperature sensors, and text and audio input devices.
- *Grasping and Manipulation*: This domain refers to a robot's capacity to handle physical objects and perform actions that modify the environment. Essential sensors and hardware for grasping and manipulation include suction cups, grippers, tails, arms, motors, and tactile sensors.
- *Communication with Humans and Other Robots*: Robots can possess the ability to interact with humans and coordinate with other homogeneous or heterogeneous robots.

4.1 Challenges and Opportunities in Locomotion

While not all robots require locomotion capabilities, many do (as shown in Table 1). For instance, robots like *PARO*, *Da Vinci*, and *Moley*, which provide therapeutic, surgical, or culinary assistance, respectively, are anchored to a specific platform or surface. Their motion is limited to the range of their manipulators, denying them independent mobility. However, a plethora of assistive applications necessitates the ability of a robot to move autonomously or

ID	Robot Name	Appearance	Locomotion	Perception	Grasping & Manipulation	Communication with Human	Communication with Robots	Local User	Tele-operator /tele-user	Bystander	AI Developer	Purpose
1	Astro	HM	1	1	0	1	0	1	1	1	0	Household, service, social
2	Atlas	H	1	1	1	1	0	1	1	1	1	Industrial
3	Ava	HM	1	1	0	1	0	0	1	1	0	Telepresence, household, service
4	Baxter/Sawyer	HM	1	1	1	1	0	1	1	1	1	Industrial/Research and education
5	Bluerov	M	1	1	0	0	0	0	0	1	1	Underwater navigation
6	Da Vinci	M	0	1	1	1	0	0	1	1	0	Surgical robot
7	DJI drone	M	1	1	0	0	0	0	0	1	1	Aerial navigation
8	EMO	HM	1	1	0	1	0	1	0	1	0	Small pet, social
9	Everyday robot	HM	1	1	1	1	0	1	1	1	0	Service (discontinued)
10	Fetch	HM	1	1	1	1	0	1	1	1	1	Research and education
11	Franka	M	0	1	1	1	0	1	1	1	1	Research and education
12	Kilobot	A	1	1	0	1	1	0	0	0	1	Swarms of robots
13	LoCoBot	M	1	1	1	1	0	1	1	1	1	Research and education
14	Moley	HM	0	1	1	1	0	1	0	0	0	Household, service, food prep
15	NAO	H	1	1	1	1	0	1	1	1	1	Social
16	Ohmni	HM	1	1	0	1	0	0	1	1	0	Telepresence
17	PARO	A	0	1	0	1	0	1	0	0	0	Social, therapeutic pet
18	Pepper	H	1	1	1	1	0	1	1	1	1	Social
19	Phoenix	H	1	1	1	1	0	0	1	1	1	Service, tele-operated
20	Plato	HM	1	1	0	1	0	1	1	1	0	Service, food delivery, social
21	PR2	HM	1	1	1	1	0	1	1	1	1	Household, service, industrial
22	Reachy	HM	1	1	1	1	0	0	1	1	1	Telepresence, service, social
23	Roomba	M	1	1	1	1	0	1	1	0	0	Household, service
24	Relay	HM	1	1	0	1	0	1	0	1	0	Service, hospitality
25	Spot	A	1	1	1	1	0	1	1	1	1	Service, industrial, dog-like
26	Stretch	M	1	1	1	1	0	1	1	1	1	Industrial
27	TurtleBot	M	1	1	1	1	0	1	1	1	1	Research and education
28	Unitree Go 1	A	1	1	1	1	0	1	1	1	1	Research and education
29	Vgo	HM	1	0	0	1	0	0	1	1	0	Telepresence

Table 1. A sample of robots classified according to the proposed robot appearance dimension. “H” stands for “Human-like”, “A” stands for “Animal-like”, “M” stands for “Machine-like”, and “HM” stands for “Human- and Machine like”.

semi-autonomously from one location to another, a capability known as locomotion [135]. Robotic systems often take cues from biological locomotion mechanisms, adopting strategies such as crawling, sliding, running, jumping, or walking. The majority of locomotive robots employ three types of hardware: i) wheeled robots (*Reachy, Everyday robot,*

Relay, Pepper, NAO, Pluto, Astro, Turtlebot, LoCoBot, Roomba, Ohmni); ii) chain or track systems; and iii) legged robots (*Atlas, Phoenix, Spot, Unitree Go 1*). The design of the mechanics, materials, and degrees of freedom of this hardware pose significant challenges. Software design challenges for enabling locomotion include addressing high-level robotic tasks such as simultaneous localization and mapping (SLAM) and visual navigation, along with the associated issues these areas present.

4.1.1 ***Simultaneous Localization and Mapping (SLAM).*** In the SLAM task, a robot must concurrently reconstruct its environment and determine its position (rotation and translation) relative to a global coordinate frame. Various types of SLAM tasks, including Visual SLAM and Semantic SLAM, face the common challenges of accurate feature extraction, loop closure, and real-time computation [110, 143]. These tasks require substantial computational resources and rich sensor data, which can be prohibitive in terms of cost for many robots.

In our robot-mediated assistance context, an *opportunity* arises: the robot could request the bystanders or tele-operators to supply a coarse map or landmarks, thereby simplifying the map-building process and reducing computing resource demands. For instance, users could sketch a map on their smartphone app for the cleaning robot, *Roomba*. An additional *opportunity* arises in the realm of tele-operation, where the objective is to localize the robot under low-bandwidth communication channels. In such circumstances, rather than streaming the full video feed from the robot's camera, which requires a higher bandwidth, the robot could opt to transmit only a select few anchor points, specifically, image feature descriptors, grounded in visual-inertial odometry [41].

4.1.2 ***Visual Navigation.*** Traditional visual navigation encompasses three fundamental stages: *i) mapping:* the process of generating a representation of the environment, capturing relevant features and landmarks; *ii) localization:* the robot identifies its position within the mapped environment, typically through sensor data comparison with mapped features; *iii) path planning:* given the mapped environment and the robot's current location, an optimal path is determined to reach the desired destination [114]. More recently, datasets like Habitat 2.0 [140] offer simulation platforms that enable the exploration of unseen 3D environments for benchmarking two versions of visual navigation tasks: image-goal navigation and object-goal navigation. In the context of image-goal navigation, the objective for a robot is to autonomously reach a specific goal indicated by an image representing the target location. This task requires the robot to interpret the visual information provided by the image and make decisions on how to navigate the environment to successfully reach the desired destination [44, 109, 166]. Visual navigation presents numerous challenges, such as addressing variations in lighting conditions, adapting to dynamic environments, and maintaining real-time processing capabilities.

An opportunity exists to design robots that can navigate within an environment semi-autonomously, using input or cues from providers (i.e., non-user humans) [161]. For example, the robot could follow a human traveling to the same destination. This approach could also assist the robot in recovering from navigational failures.

4.2 Challenges and Opportunities in Perception

Perception in robotics hinges on the extraction of meaningful data from the robot's array of sensors. Modern robots come equipped with a variety of sensors, such as visual cameras, thermal sensors, infrared sensors, and LiDAR (see Section 4). Of these, the visual camera emerges as the most impactful sensor due to its prevalent use and cost-effectiveness on a large scale. Notably, the evolution of machine learning, specifically deep learning techniques, has pushed computer vision methods to a stage where they can reliably provide solutions for robot-mediated assistance. Despite these positive developments, challenges persist. The system must handle visual data to tackle a range of computer vision tasks. These

tasks span basic, mid-level, and high-level functions, such as object recognition, object detection, object tracking, segmentation, scene comprehension, and 3D reconstruction. In the following sections, we explore these challenges and opportunities in the context of robot-mediated assistance, considering the computer vision aspect, the robot assisting the human user and the provider.

4.2.1 ***Object Recognition.*** Object recognition involves classifying an image into specific categories using distinct feature descriptors. In the past decade, significant strides have been made in the field, predominantly through the application of Convolutional Neural Networks (CNNs) [137] and Transformers [33]. Groundbreaking networks like AlexNet [75], VGG [137], and ResNet [54] have been pivotal in driving these advancements. Despite progress, challenges persist in object recognition, such as accommodating objects with differing appearances, dealing with occlusions, navigating background clutter, and adjusting to changes in orientation and perspective. Robust recognition, capable of withstanding variations in lighting conditions [61], is essential for a robot to interact effectively with its environment. Domain adaptation strategies provide potential solutions to these challenges, particularly when dealing with variations in appearance, lighting conditions, and object distribution [43, 124, 138].

There lies an *opportunity* to apply active learning [65] in object recognition. Users can identify and collect instances where they observe the robot's failures in recognition scenarios, and they can provide the correct annotations. This user-provided information can subsequently improve the model. Additionally, there is a potential *opportunity* for the development of novel techniques specifically designed to recognize content on digital displays [81]. This difficulty arises from the factors such as varying screen brightness, the mingling of different light sources (e.g., LCD backlight, sunlight, lamplight), and mismatches in frame rates between the camera and the screen — all can influence the robot's ability to accurately interpret digital content. Further complications arise from inconsistencies between the pixel grid dimensions of the camera and the digital screen, leading to moiré patterns, which manifest as strobe or striping optical effects[115].

4.2.2 ***Object Detection.*** Object detection involves the classification of object categories within an image and determining their positions, typically depicted by rectangular bounding boxes. Challenges in object detection span various factors such as diverse object appearances, occlusions, background clutter, variations in object pose, and detection of object parts [38]. For robot-mediated assistance, it is vital for the robot to accurately detect the objects with which it interacts [18, 53]. This requirement becomes especially significant when the robot needs to grasp or manipulate objects.

This arises an opportunity for applying a human-in-the-loop technique [17]: the human provider can assist the robot in detecting objects in cases where it encounters difficulties; then they can proactively identify challenging scenarios and capture images of those situations. These challenging images can then be utilized to train a robust supervised model through a combination of online learning and interactive labeling, even in the presence of weak supervision.

4.2.3 ***Object Tracking.*** Object tracking refers to the continuous detection of an object within a spatio-temporal setting, such as a video. Object tracking faces numerous challenges, such as illumination variations, abrupt camera motion, motion blur, fast motion [55], rotation of the target object (within or outside the image plane), object deformation, occlusion (partial or complete), the target object going out of view, changes in viewpoint, scale variations, low-resolution images, background clutter, and distinguishing between similar objects [74]. The robot might be required to track the human user's movements. Moreover, there may be scenarios where the robot needs to track a moving target [45].

An opportunity is to develop a single object tracker [36] to continuously track a unique individual based on their non-facial features (e.g., heights, body shapes, and gaits). An extension of this tracker is to track multiple objects

based on natural language commands (e.g., "keep an eye on Tom, the cat, and Jerry, the mouse") [42, 152]. The second opportunity is to improve the trackers by robust feature learning [156].

4.2.4 ***Segmentation.*** Segmentation involves partitioning an image into distinct regions based on particular grouping criteria. The main challenge in segmentation stems from the diverse and intricate nature of environments, including indoor [136], outdoor [3], aerial [100], and underwater settings [59]. Various types of segmentation, such as semantic segmentation [8, 24, 94], instance segmentation [32, 123], and panoptic segmentation [70], are employed depending on the specific application needs. A significant challenge in semantic segmentation is the inconsistencies found in labeling across different benchmarks [79].

Several opportunities present themselves in the realm of robot-mediated assistance. First, there is the potential to create a unified model for segmentation [26]. Second, it is possible to design segmentation models with robust zero-shot generalization capability, which can effectively segment any object found in diverse environments [71]. The third opportunity involves creating segmentation models that can respond to natural language prompts (e.g., *"segment the brown dog in the image"*). This concept of *promptable segmentation* aligns with challenges in visual question answering (VQA) [48], but with a focus on segmentation. Finally, there is a promising opportunity to develop optical character recognition (OCR) techniques that can adeptly segment text on curved or warped surfaces, like reading text from a round medicine bottle or a crumpled piece of paper.

4.2.5 ***Scene Understanding.*** Scene understanding aims to build a system with the capacity to fully perceive and interpret a scene, taking into account its layout, components, and environmental characteristics. This task is particularly challenging due to the complexities and diversity inherent in different environments; the robot should be capable of answering a multitude of queries, such as identifying the labels and locations of various regions, discerning the spatial extent of objects, understanding the layout of the scene, and recognizing different scene types, like a kitchen, office, or bedroom [91, 128, 151, 159].

A promising avenue to explore is the development of the robot's ability to detect changes within dynamic environments alongside the human user and other active or passive bystanders [1]. These could include changes in the number of objects, either added or removed, as well as shifts in the lighting ambiance.

4.2.6 ***Visual Question Answering.*** Visual Question Answering (VQA) presents the challenge of predicting an open-ended answer to a question given an image [7]. Swift advancements in computer vision and natural language processing (NLP) methodologies, along with the accessibility of large-scale datasets, have sparked significant interest in VQA [64]. There is a wide range of deep neural architectural choices used in the design of existing VQA models, including attention mechanisms [97], stacks of attention networks [158], combinations of LSTM and CNN modules [96], compositional models [58], and transformer-based attention networks [25, 51, 87, 95, 141]. Successful VQA models hold great promise for robot-mediated assistance as they allow users to pose natural language questions to robots on a variety of topics, such as visual recognition, reasoning, object counting, and common sense. A notable challenge in Visual Question Answering (VQA) models centers on their robustness, i.e., their capacity to reliably respond to slightly varied forms of the same question [88] such as *linguistic shifts, logical reasoning adjustments, visual content manipulation,* and *changes in answer distribution between training and testing datasets*. Current strategies to boost robustness revolve around crafting purpose-designed benchmarks and robustness metrics [5, 49, 66, 130].

A crucial opportunity within robot-mediated assistance is the generation of a more sophisticated context for each question, leveraging other computer vision tasks, particularly scene understanding and real-time segmentation. Such

enriched contextual understanding can contribute substantially to the accuracy and applicability of the VQA models in complex real-world scenarios.

4.2.7 ***3D Scene Reconstruction.*** Several approaches can be taken for 3D scene reconstruction, depending on the availability of cameras and their relative calibration. For instance, *monocular depth estimation* is a technique that aims to predict the depth of each pixel from a single image [12, 35, 125]. *Stereo matching* estimates the disparity between corresponding pixels in a pair of images taken by a binocular camera setup. This process determines the displacement value for each pixel, providing vital depth information about the scene [90, 127, 155]. Finally, *multiview stereo* attempts to 3D reconstruct a scene using multiple images. These images can be taken by a single camera from various viewpoints or multiple calibrated cameras [40, 129, 160]. There are several challenges associated with 3D reconstruction, including the handling of textureless regions, self-occlusion, object occlusion, deformable objects, dynamic scenes with moving objects, low-light or nighttime scenarios, and the requirement for real-time processing [50, 68, 77, 165]. These challenges pose obstacles in accurately reconstructing the three-dimensional structure of the scene and require innovative solutions to address them.

Considerable improvements to these challenges would create several opportunities for robot-mediated assistance. Firstly, effective 3D scene reconstruction can significantly improve the robot's spatial awareness, leading to a more comprehensive understanding of its surroundings. Secondly, 3D object recognition can enhance the robot's ability to interact with its environment, particularly in terms of object manipulation and grasping.

4.3 Challenges and Opportunities in Grasping and Manipulation

Robotic grasping and manipulation are essential in unstructured environments such as homes, offices, warehouses, and logistics centers where there is a physical necessity to retrieve objects. Two key representations for grasp configurations exist: i) planar grasping, which utilizes a simpler representation (a 2D location) but is limited in its applicability [84, 99], and ii) 6-DOF (Degrees of Freedom) grasping, which is more complex but provides a greater range of flexibility [108].

Grasping challenges. Grasping is challenging because the task of grasp synthesis, i.e., finding a grasp configuration for a given object, involves determining the most suitable configuration from a wide range of plausible candidates that meet specific criteria [16]. The grasping problem can be approached as an object detection problem that involves predicting the rectangular location on the object that can be grasped [84]. Large-scale grasp datasets are typically built by sampling from a simulated environment, using an analytical method [99, 108] or self-supervision [85, 118]. The challenge of grasping becomes even more pronounced when the target object is occluded by clutter, a scenario framed as *mechanical search* [31, 76]. Common mechanical search strategies involve a series of actions, including parallel jaw grasping, suction grasping, and pushing, until the target object is successfully retrieved [31].

Manipulation challenges. Dexterous manipulation tasks also remain challenging, including actions such as *throwing, sliding, poking, pushing* in unclutter environments, *stacking boxes according to their sizes/weights*, to more complex tasks such as *inserting pegs in holes, hammering, sorting objects, packaging objects, folding clothes, pouring fluid into a glass.* Additionally, challenges arise in manipulating deformable objects (e.g., clothes), performing actions that cause changes to the object (such as cutting or crushing), and executing in-hand manipulations where an object is moved while being held (e.g., twirling a pen between fingers). While robots can learn manipulation tasks through human demonstrations with fewer trials, these methods often struggle with generalization [14].

Opportunities. Opportunities exist in developing robots capable of universally picking objects of different sizes and shapes. Improvements in these areas could involve developing more efficient algorithms or enhancing the motion capabilities of robotic hardware. Dexterous manipulation presents open opportunities, specifically in developing flexible gripping mechanisms that require research in new hardware and materials for effective grasping. Indeed, in robot-mediated assistance, the ability of a robot to adapt and use tools or utensils appropriately presents a significant opportunity. This requires the development of adaptive algorithms that can teach robots how and when to use these tools effectively. For instance, consider a scenario where a robot needs to pick up a small object like an M&M chocolate. In this case, it may be more effective for the robot to use kitchen clamps or tongs instead of trying to grip the small piece with its own mechanical fingers or suction apparatus. This adaptive tool use can make robotic assistance more practical and efficient in various real-world situations. It also presents an opportunity for robots to engage in more complex tasks that would traditionally be challenging for them. Adapting to use tools also means that the robot needs to understand the properties and functionality of different tools, which might involve a combination of techniques from computer vision, machine learning, and reinforcement learning. It can also entail the robot learning from past experiences, much like how humans learn to use tools effectively over time. Further discussions on challenges and opportunities can be found in works such as [14, 28].

4.4 Challenges and Opportunities in Human-Robot Interaction

Recent developments in robotics have seen a transition from stationary, solitary robotic systems to emerging categories of mobile manipulators. These new robot types are intended to operate within human-centric environments and work in direct cooperation with people. They are also accommodating a wide spectrum of user profiles that span diverse backgrounds, physical and cognitive capabilities, training, and receptivity to technological adaptation. As such, the significance of Human-Robot Interaction (HRI) and specifically, a robot's capability to communicate with humans, has been amplified. Despite the considerable progress made in HRI over the past decade, this field continues to face several significant challenges and opportunities. A few key ones are discussed below:

4.4.1 ***Promoting Sustained Interaction, Mitigating Overreliance, and Ensuring Cultural Sensitivity.***

Sustained interaction. Ensuring substantive, long-term interaction between robots and users represents a crucial challenge. Within HRI literature, "long-term" is commonly understood to mean a period of approximately two months [67, 139], which is typically sufficient for a user to become familiar with a robot, thus neutralizing novelty bias [83].

Various social robots, such as PARO [149] and Jibo [116], provide immediate comfort or diversion but struggle to foster lasting connections or adapt their behaviors based on past user interactions. For example, an ideal in-home robot caregiver should adjust its care strategies over time according to the resident's preferences, habits, and health transitions. However, most social robots presently serve merely as conversation partners, limited to basic dialogues [83]. Similarly, while assistive robotic systems like exoskeletons and rehabilitation robots can provide physical support, their adaptability to individual user needs and preferences remains insufficient [98].

Mitigating overreliance. Conversely, an essential challenge is the potential for users to become overly dependent on robots. As humans tend to treat anthropomorphic technologies as social entities [113, 122, 131], there's a risk that strong social bonds with robots might replace meaningful human relationships, for instance, in the context of elderly care [146]. This dynamic also raises concerns about possible exploitation via anthropomorphic robots.

Cultural sensitivity. The necessity of cross-cultural interaction and cultural sensitivity is equally significant for successful HRI. Robots such as Pepper can recognize and respond to a variety of human emotions; however, cultural differences in expressing emotions could lead to misunderstandings. For instance, excessive direct eye contact, often seen as disrespectful in Japan, is typically associated with honesty and engagement in Western cultures. Thus, a robot designed based on Western norms might inadvertently offend users from non-Western cultures. The development of culturally adaptable robots, therefore, constitutes an essential component of HRI research [11].

Consequently, fostering personalization, maintaining long-term adaptability, striking a balance between human and robot interaction, and ensuring cultural appropriateness, continue to be active areas of investigation within HRI [83].

4.4.2 ***Robots' Understanding of Humans.*** Interpreting and understanding human behavior remains a major challenge in the areas of AI and HRI [10, 19]. This problem encompasses the ability to perceive and interpret human non-verbal cues, gestures, and actions, as well as learn to emulate their behaviors, strategies, and actions.

Interpreting Non-verbal Cues. Non-verbal communication is intricate, often representing a significant stumbling block for robots. For example, the Kismet robot, designed for social interaction [20], can recognize and respond to a certain range of predefined facial expressions and vocal cues. However, it generally falls short in dealing with the subtleties and diversity intrinsic to human non-verbal communication. Humans use body language, facial expressions, voice modulation, and even periods of silence to convey complex emotions and intentions. The task of equipping robots with the ability to decipher these non-verbal cues and respond appropriately remains a central challenge in AI and HRI [10].

Gestural Recognition. Recognition of gestures, including hand movements and body language, can enable more intuitive, non-verbal interaction between humans and robots. The inherent challenges associated with gestural recognition involve distinguishing ambiguous gestures, managing occlusions, and handling the complexity of human body movements [9, 92, 153, 162]. It also includes real-time gestural classification with minimal latency [103].

Learning from Human Interactions. Facilitating the acquisition of complex, long-term domestic tasks by robots also remains an active area of research. Presently, most assistive robots function under open-loop paradigms [111]—performing tasks based on explicit human commands—either locally or remotely [29], or adhering to pre-programmed routines. Some research attempts to enable humans to guide robots through demonstrations or performance feedback. However, providing exhaustive feedback can prove taxing for users, particularly when it requires comprehensive task demonstrations or intricate instruction. Certain feedback mechanisms may not be feasible, especially when preferences only emerge in response to specific stimuli. Furthermore, swiftly adapting robots to new partners is a challenging learning problem, given that partners may have varying preferences, capabilities, knowledge, intentions, or strategies for accomplishing shared objectives [46, 107, 164].

Various strategies are proposed to address these challenges. For example, one approach involves modeling humans as partially observable Markov decision processes (POMDPs) [63, 78, 104] in an adaptation parameter space, which a robot can learn and apply to direct its actions. By learning the human's adaptation parameter [164], the robot can modify its behavior to foster more effective collaboration [72]. This modification can be assessed using metrics such as task performance, collaborative fluency [57], team trust [52, 86], or engagement [134]. Recent advances also utilize deep neural networks, including large language models [6], to help robots incorporate feedback from natural user behavior, framing the learning problem as an online Inverse Reinforcement Learning (IRL) process [2, 111].

4.4.3 ***Security, Privacy, Trust, and Ethical Considerations.*** With the increasing integration of robots into human society, the urgency of addressing security, privacy, and ethical concerns is magnified. Consider, for instance, a caregiving robot designed to alleviate the physical strain on human caregivers by assisting with lifting and moving patients. This scenario prompts ethical questions surrounding patient consent.

Security Concerns. Three primary categories emerge in relation to robots and security and privacy issues: i) facilitation of surveillance, as demonstrated by police drones overseeing public protests; ii) infringement upon traditionally private domains, such as home robots that fall victim to hackers; and iii) creation of a perception of being watched, a prevalent concern associated with anthropomorphic (human-like) robots [21, 28]. Moreover, the possibility of security vulnerabilities can increase the potential for physical harm that a compromised robot could cause.

Privacy Challenges. Ensuring visual privacy presents another hurdle, as it necessitates the anonymization or obscuration of identifiable characteristics without impairing the accuracy of other vision tasks [150]. Determining which information qualifies as "identifiable" or "private" contributes to this complexity. Critically, the algorithms propelling these vision tasks should not direct users toward corporate or other objectives conflicting with their own interests [112]. Hence, a balance between the beneficial and detrimental impacts of robots on humans mandates thorough research and mindful consideration within the realm of HRI.

Trust in HRI. Trust forms a vital component of successful HRI. Users must understand a robot's abilities and intentions and feel a sufficient level of control over the robot's actions [52]. Trust extends beyond merely the user's faith in the robot, encompassing the robot's estimation of the user's trust as well. While there is prior work on the factors influencing trust in robots, constructing robust models to manage trust proves to be a challenge [86, 126]. For instance, while consistency and reliability may be desired characteristics, unexpected behavior might be appreciated in certain contexts due to its novelty, such as a robot cheating during gameplay to maximize entertainment [133]. Striking a balance between trust, reliability, and novelty, therefore, presents an area for exploration.

Rights of Robots. One of the significant ethical issues in robot-mediated assistance concerns the rights of robots against abusive behavior, such as wanton kicking of a robot for amusement. This concern also prompts questions about the ramifications of assigning rights to robots. A contrasting viewpoint proposes considering robots as bearers of 'rites' instead of 'rights' [69]. 'Rites' denote a sequence of actions, often involving multiple actors, such as robots, users, and providers, collectively bearing symbolic significance. Through these rites, actors acknowledge the value of the interaction and delineate their reciprocal roles. For instance, in a basketball game, each player has specific roles or duties, thus, their rites. Should player 1 neglect to pass the ball to a better-positioned teammate, player 2, player 1 is not infringing upon player 2's rights but failing in their own rites. Enhancing our comprehension of the nuanced roles, rites, and rights affiliated with robots offers another promising area for investigation within HRI.

4.4.4 ***Lack of Open Robot Platforms, Datasets, and Real-World Evaluation Environments.*** The spectrum of human-robot interactions is extensive and poses considerable challenges in objective evaluation [28]. This spectrum spans from scenarios where humans have complete control over robots to those where robots operate with minimal human oversight and even scenarios where robots are engaged in social interactions.

The first challenge lies in the scarcity of open robot platforms suitable for a wide range of HRI research. The majority of research platforms are primarily designed for mobility in 2D or 3D space (e.g., ground, air), but effective interaction with humans demands careful consideration of a robot's appearance (form) and its perceptual and operational capabilities

(function). We note that middleware such as the Robot Operating System (ROS) has spurred considerable adoption in the field. Affordably priced humanoid robots, expressive enough for social HRI research, are not available at present. For instance, *Pepper* robot, despite lacking facial features, is still priced well above a thousand dollars. Some startups occasionally introduce potentially promising platforms (e.g., Jibo, Kuri), but these offerings tend to be short-lived. Further, there is a dearth of economically priced tabletop robot platforms designed for social and socially assistive HRI. These platforms would aid in the creation, deployment, and testing of large-scale user studies.

The second challenge is the lack of large-scale, real-world datasets for human-robot interaction. Advancements in HRI for specific user groups, such as the elderly, children, stroke patients, or individuals on the autism spectrum, are contingent on the availability of interaction data with these populations. However, the collection and dissemination of such data are hindered by substantial privacy constraints; consequently, only a limited number of researchers are able to conduct extensive studies with real-world datasets. A few community datasets and testbeds are available at present, such as the Minedojo [37], a Minecraft-based simulated environment for developing and testing reinforcement learning algorithms. Other platforms include Overcooked multi-agent environment [22], IKEA furniture assembly environment [82], and scalable data collection platform via teleoperation [117]. However, these simulations, while useful, are still different from real-world interactions.

Lastly, there is a lack of access to real-world evaluation settings, such as nursing homes, schools, and hospitals. This restriction often compels HRI researchers to rely on university students as study participants, potentially skewing the results due to this biased sample. Mirroring the impact of large-scale datasets in computer vision research, creating shared resources, datasets, and testbeds presents an open opportunity in HRI that can allow for methodical comparisons, fostering collaboration and cumulative progress across the field.

4.4.5 ***Novel Interaction Modalities in HRI.*** HRI brings fresh opportunities to interaction design, extending beyond the traditional scope of human-computer interaction. Traditionally, the human-computer interaction paradigm is rooted in the windows-icons-menu-pointer (WIMP) model [148]. In this paradigm, users interact with application windows, menus, and visual metaphors (icons) using pointing devices like a mouse. However, HRI offers the potential to revolutionize this model with innovative interaction modalities. This includes the integration of virtual and augmented reality (VR/AR) technologies, enabling shared viewpoints, contexts, and embodied interactions between humans and robots. These advancements promise a wealth of design possibilities for various input devices, the development of new interaction models, and the production of reference implementations for hardware and software layers.

Immersive Interfaces in Virtual Reality (VR). VR-based immersive interfaces are emerging as pivotal tools for teleoperated robots such as *Reachy*, enabling human users to remotely supervise and control multiple robots [47]. This interface can facilitate collaboration and mimic the spontaneity found in face-to-face interactions. For instance, robots possessing movable necks can mimic human gestures like nodding or adjusting their gaze direction, which amplifies the naturalness of interactions. Furthermore, the user's avatar in the VR environment can perform actions such as shaking hands with bystander avatars or even engaging in tasks like cleaning a kitchen countertop located in a remote environment. These dynamic, embodied interactions can substantially boost user engagement, cultivating a sensation of true presence in the virtual scene.

This novel interface also presents opportunities to minimize interaction latency by improving rendering pipelines and utilizing progressive compression in the communication channels. Additional opportunities include developing technologies that can accurately create realistic 3D environments with less manual effort, such as generative AI models

that produce images from textual prompts. This development aims to enhance not just visual realism but also interaction realism, pertaining to how the robot responds to user commands or behaves within the VR environment.

Brain-Control Interfaces. Brain-control interfaces (BCIs) have driven substantial advancements in HRI, notably allowing individuals with quadriplegia to operate robotic arms via thought processes alone [80, 142]. These interfaces can translate brain activity into control signals, establishing a new communication pathway between humans and robots. Despite significant progress, research efforts are focused on enhancing the reliability, responsiveness, and versatility of BCIs. Reliability ensures consistent operation, a critical requirement in healthcare or assistive technologies. Response time improvement facilitates smoother and more timely interactions, which is crucial in emergency or surgical applications. Lastly, improving adaptability encompasses making the interface user-friendly and capable of catering to various users and conditions. Such advances are poised to revolutionize human-robot interaction.

Collaborative Robot Interfaces. Collaborative robots, often referred to as "cobots," represent a significant development in Human-Robot Interaction (HRI) [4]. These machines are engineered with a host of advanced sensors, actuators, and control algorithms to ensure both safe and efficient interactions with human colleagues in shared work environments.

Despite the potential benefits, several challenges lie ahead for the successful integration of cobots into a wide variety of work settings. Firstly, refining their perception capabilities and decision-making processes is vital to enhance their awareness and responsiveness towards human co-workers. Secondly, managing the change in workflow that comes with introducing cobots into traditional work environments is a significant hurdle. Lastly, there is a need for rigorous, context-specific evaluations of their performance, especially in scenarios requiring close collaboration between humans and cobots. Addressing these challenges is critical to ensure seamless and intuitive human-cobot communication, thereby enhancing the efficacy of tasks that require shared effort and understanding.

5 CONCLUSION

In this work, we present a nuanced analysis of robot-mediated assistance by introducing three new lenses to extend the current framework [157]. We delve into robot appearances, explore diverse human and entity roles in human-robot interaction, and probe the spectrum of necessary robotic abilities for effective human assistance. Our in-depth analysis is derived from a comprehensive review of 29 robots, identifying pivotal challenges and untapped research avenues in computer vision and HRI within the context of robot-mediated assistance. This enriched understanding illuminates precise directions for future exploration. We anticipate our contributions will inspire multidisciplinary research communities, propelling advancements in robot-mediated assistance.

REFERENCES

[1] 2023. https://nikosuenderhauf.github.io/roboticvisionchallenges/cvpr2023. In *Workshop in IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[2] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 1.

[3] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. 2018. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision* 126, 9 (2018), 961–972.

[4] A Adriaensen, F Costantino, G Di Gravio, and R Patriarca. 2022. Teaming with industrial cobots: A socio-technical perspective on safety analysis. *Human Factors and Ergonomics in Manufacturing & Service Industries* 32, 2 (2022), 173–198.

[5] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.

[6] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on computer vision*.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla. 2017. SegNet: a deep convolutional encloder-decoder architecture for scene segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[9] Gilles Bailly, Jörg Müller, Michael Rohs, Daniel Wigdor, and Sven Kratz. 2012. Shoesense: a new perspective on gestural interaction and wearable applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1239–1248.

[10] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. 2017. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence* 242 (2017), 132–171.

[11] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[12] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4009–4018.

[13] Aaron Biggs. 2015. Paro robot. https://commons.wikimedia.org/wiki/File:Paro_robot.jpg. Licensed under CC BY-SA 2.0 via Wikimedia Commons.

[14] Aude Billard and Danica Kragic. 2019. Trends and challenges in robot manipulation. *Science* 364, 6446 (2019), eaat8414.

[15] Mike Blow, Kerstin Dautenhahn, Andrew Appleby, Chrystopher L Nehaniv, and David Lee. 2006. The art of designing robot faces: Dimensions for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 331–332.

[16] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. 2013. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics* 30, 2 (2013), 289–309.

[17] Steve Branson, Pietro Perona, and Serge Belongie. 2011. Strong supervision from weak annotation: Interactive training of deformable part models. In *2011 International Conference on Computer Vision*. IEEE, 1832–1839.

[18] Garrick Brazil, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. 2023. Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[19] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International journal of human-computer studies* 59, 1-2 (2003), 119–155.

[20] Cynthia Breazeal. 2004. *Designing sociable robots*. MIT press.

[21] Ryan Calo. 2010. Open robotics. *Md. L. Rev.* 70 (2010), 571.

[22] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32 (2019).

[23] Wan-Ling Chang and Selma Šabanović. 2015. Interaction expands function: Social shaping of the therapeutic robot PARO in a nursing home. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 343–350.

[24] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

[26] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Michael A Chilton, Bill C Hardgrave, and Deborah J Armstrong. 2010. Performance and strain levels of it workers engaged in rapidly changing environments: a person-job fit perspective. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 41, 1 (2010), 8–35.

[28] Henrik Christensen, Nancy Amato, Holly Yanco, Maja Mataric, Howie Choset, Ann Drobnis, Ken Goldberg, Jessy Grizzle, Gregory Hager, John Hollerbach, Seth Hutchinson, Venkat Krovi, Daniel Lee, Bill Smart, Jeff Trinkle, and Gaurav Sukhatme. 2021. A Roadmap for US Robotics - From Internet to Robotics 2020 Edition. *Foundations and Trends in Robotics* 8, 4 (2021), 307–424. https://doi.org/10.1561/2300000066

[29] Matei Ciocarlie, Kaijen Hsiao, Adam Leeper, and David Gossow. 2012. Mobile manipulation through an assistive home robot. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5313–5320.

[30] H.H. Clark, H.H. Clark, American Council of Learned Societies, H.C. Clark, and H.H. Clark. 1996. *Using Language*. Cambridge University Press. https://books.google.ie/books?id=DiWBGOP-YnoC

[31] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Goldberg. 2019. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 1614–1621.

[32] Bert De Brabandere, Davy Neven, and Luc Van Gool. 2017. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551* (2017).

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[34] John V Draper, David B Kaber, and John M Usher. 1998. Telepresence. *Human factors* 40, 3 (1998), 354–375.

[35] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

[36] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5374–5383.

[37] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=rc8o_j8I8PX

[38] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*. Ieee, 1–8.

[39] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[40] Yasutaka Furukawa and Jean Ponce. 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence* 32, 8 (2009), 1362–1376.

[41] Giovanni Fusco and James M Coughlan. 2018. Indoor localization using computer vision and visual-inertial odometry. In *International Conference on Computers Helping People with Special Needs*. Springer, 86–93.

[42] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4340–4349.

[43] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[44] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. 2022. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15460–15470.

[45] Samuel Gibbs. 2021. Amazon launches home robot Astro and giant Alexa display. *The Guardian* (2021).

[46] Matthew Craig Gombolay, Cindy Huang, and Julie Shah. 2015. Coordination of human-robot teaming with human task preferences. In *2015 AAAI Fall Symposium Series*.

[47] Michael A Goodrich, Jacob W Crandall, and Emilia Barakova. 2013. Teleoperation and beyond for assistive humanoid robots. *Reviews of Human factors and ergonomics* 9, 1 (2013), 175–226.

[48] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[50] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 12 (2020), 4338–4364.

[51] Tanmay Gupta, A. Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2021. Towards General Purpose Vision Systems. *ArXiv* abs/2104.00743 (2021).

[52] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[53] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[55] Jo ao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 583–591.

[56] Wan Ching Ho, Kerstin Dautenhahn, Mei Yii Lim, Patricia A Vargas, Ruth Aylett, and Sibylle Enz. 2009. An initial memory model for virtual and robot companions supporting migration and long-term interaction. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 277–284.

[57] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[58] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*. 53–69.

[59] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. 2020. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1769–1776.

[60] David Jaffe, John Thiemer, and Drew Nelson. 2012. Perspectives in Assistive Technology. https://web.stanford.edu/class/engr110/2012/04b-Jaffe.pdf

[61] Anne Jorstad, David Jacobs, and Alain Trouvé. 2011. A deformation and lighting insensitive metric for face recognition based on dense correspondences. In *CVPR 2011*. IEEE, 2353–2360.

[62] Steve Jurvetson. 2013. Baxter Robot Caught Coding. https://commons.wikimedia.org/wiki/File:Caught_Coding_(9690512888).jpg. Licensed under CC BY 2.0 via Wikimedia Commons.

[63] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[64] Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding* 163 (2017), 3–20.

[65] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. 2007. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th international conference on computer vision*. IEEE, 1–8.

[66] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2776–2785.

[67] Cory D Kidd and Cynthia Breazeal. 2008. Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3230–3235.

[68] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. 2018. Multispectral transfer network: Unsupervised depth estimation for all-day vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[69] Tae Wan Kim and Alan Strudler. 2023. Should Robots Have Rights or Rites? *Commun. ACM* 66, 6 (2023), 78–85.

[70] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9404–9413.

[71] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[72] Kheng Lee Koay, Dag Sverre Syrdal, Michael L Walters, and Kerstin Dautenhahn. 2007. Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 564–569.

[73] Hema S Koppula, Ashesh Jain, and Ashutosh Saxena. 2016. Anticipatory planning for human-robot teams. In *Experimental Robotics: The 14th International Symposium on Experimental Robotics*. Springer, 453–470.

[74] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Hyung Jin Chang, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Johanna Bjorklund, Yushan Zhang, Zhongqun Zhang, Song Yan, Wenyan Yang, Dingding Cai, Christoph Mayer, and Gustavo Fernandez. 2022. The Tenth Visual Object Tracking VOT2022 Challenge Results.

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[76] Andrey Kurenkov, Joseph Taglic, Rohun Kulkarni, Marcus Dominguez-Kuhne, Animesh Garg, Roberto Martín-Martín, and Silvio Savarese. 2020. Visuomotor mechanical search: Learning to retrieve target objects in clutter. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8408–8414.

[77] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. 2020. A survey on deep learning techniques for stereo-based depth estimation. *IEEE transactions on pattern analysis and machine intelligence* 44, 4 (2020), 1738–1764.

[78] Chi-Pang Lam and S Shankar Sastry. 2014. A POMDP framework for human-in-the-loop system. In *53rd ieee conference on decision and control*. IEEE, 6031–6036.

[79] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. 2020. MSeg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2879–2888.

[80] Mikhail A Lebedev and Miguel AL Nicolelis. 2006. Brain–machine interfaces: past, present and future. *TRENDS in Neurosciences* 29, 9 (2006), 536–546.

[81] Sooyeon Lee, Rui Yu, Jingyi Xie, Syed Masum Billah, and John M Carroll. 2022. Opportunities for human-AI collaboration in remote sighted assistance. In *27th International Conference on Intelligent User Interfaces*. 63–78.

[82] Youngwoon Lee, Edward S Hu, and Joseph J Lim. 2021. IKEA furniture assembly environment for long-horizon complex manipulation tasks. In *2021 ieee international conference on robotics and automation (icra)*. IEEE, 6343–6349.

[83] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5 (2013), 291–308.

[84] Ian Lenz, Honglak Lee, and Ashutosh Saxena. 2015. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research* 34, 4-5 (2015), 705–724.

[85] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research* 37, 4-5 (2018), 421–436.

[86] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. *Foundations of trusted autonomy* (2018), 135–159.

[87] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[88] Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673* (2020).

[89] Wanwan Li, Javier Talavera, Amilcar Gomez Samayoa, Jyh-Ming Lien, and Lap-Fai Yu. 2020. Automatic Synthesis of Virtual Wheelchair Training Scenarios. In *IEEE Virtual Reality* (Atlanta, Georgia).

[90] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision. 6197–6206.

[91] Dahua Lin, Sanja Fidler, and Raquel Urtasun. 2013. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE international conference on computer vision*. 1417–1424.

[92] Mingyu Liu, Mathieu Nancel, and Daniel Vogel. 2015. Gunslinger: Subtle arms-down mid-air interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 63–71.

[93] Manja Lohse, Frank Hegel, and Britta Wrede. 2008. Domestic applications for social robots: an online survey on the influence of appearance and capabilities. (2008).

[94] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[95] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[96] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. 2015. Deeper LSTM and normalized CNN visual question answering model. *GitHub repository* 6 (2015).

[97] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems* 29 (2016).

[98] Paweł Maciejasz, Jörg Eschweiler, Kurt Gerlach-Hahn, Arne Jansen-Troy, and Steffen Leonhardt. 2014. A survey on robotic devices for upper limb rehabilitation. *Journal of neuroengineering and rehabilitation* 11, 1 (2014), 1–29.

[99] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. 2017. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312* (2017).

[100] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. 2015. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE international conference on computer vision*. 1689–1697.

[101] Alexander Mertens, Ulrich Reiser, Benedikt Brenken, Mathias Lüdtke, Martin Hägele, Alexander Verl, Christopher Brandl, and Christopher Schlick. 2011. Assistive robots in eldercare and daily living: Automation of individual services for senior citizens. In *Intelligent Robotics and Applications: 4th International Conference, ICIRA 2011, Aachen, Germany, December 6-8, 2011, Proceedings, Part I 4*. Springer, 542–552.

[102] Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

[103] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4207–4215.

[104] George E Monahan. 1982. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. *Management science* 28, 1 (1982), 1–16.

[105] Hans Moravec. 1988. *Mind children: The future of robot and human intelligence*. Harvard University Press.

[106] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[107] Stephan J Motowildo, Walter C Borman, and Mark J Schmit. 1997. A theory of individual differences in task and contextual performance. *Human performance* 10, 2 (1997), 71–83.

[108] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 2019. 6-DOF GraspNet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2901–2910.

[109] Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. 2019. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8846–8852.

[110] Ra'ul Mur-Artal, JMM Montiel, and Juan D Tard'os. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. In *IEEE transactions on robotics*, Vol. 31. IEEE, 1147–1163.

[111] Benjamin A Newman, Christopher Jason Paxton, Kris Kitani, and Henny Admoni. 2023. Towards Online Adaptation for Autonomous Household Assistants. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 506–510.

[112] Illah Reza Nourbakhsh. 2015. *Robot futures*. Mit Press.

[113] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.

[114] Giuseppe Oriolo, Marilena Vendittelli, and Giovanni Ulivi. 1995. On-line map building and navigation for autonomous mobile robots. In *Proceedings of 1995 IEEE international conference on robotics and automation*, Vol. 3. IEEE, 2900–2906.

[115] Gerald Oster and Yasunori Nishijima. 1963. Moiré patterns. *Scientific American* 208, 5 (1963), 54–63.

[116] Anastasia K Ostrowski, Cynthia Breazeal, and Hae Won Park. 2022. Mixed-Method Long-Term Robot Usage: Older Adults' Lived Experience of Social Robots. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 33–42.

[117] Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J Lim, Stefanos Nikolaidis, et al. [n. d.]. Assisted Teleoperation for Scalable Robot Data Collection. In *CoRL 2022 Workshop on Pre-training Robot Learning*.

[118] Lerrel Pinto and Abhinav Gupta. 2016. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3406–3413.

[119] QuarkyTale. 2018. Playing with Kilobots. https://commons.wikimedia.org/wiki/File:Playing_with_Kilobots.jpg. Licensed under CC BY-SA 4.0 via Wikimedia Commons.

[120] Sarah M Rabbitt, Alan E Kazdin, and Brian Scassellati. 2015. Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. *Clinical psychology review* (2015).

[121] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. 2019. Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 1 (2019), 1–23.

[122] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10 (1996), 236605.

[123] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. 2016. Recurrent instance segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 312–329.

[124] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 213–226.

[125] Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence* 31, 5 (2008), 824–840.

[126] Kristin E Schaefer, Tracy L Sanders, Ryan E Yordon, Deborah R Billings, and Peter A Hancock. 2012. Classification of robot form: Factors predicting perceived trustworthiness. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 56. SAGE Publications Sage CA: Los Angeles, CA, 1548–1552.

[127] Daniel Scharstein and Richard Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision* 47 (2002), 7–42.

[128] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun. 2013. Box in the box: Joint 3d layout and object reasoning from single images. In *Proceedings of the IEEE International Conference on Computer Vision*. 353–360.

[129] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 1. IEEE, 519–528.

[130] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6649–6658.

[131] Amanda Sharkey and Noel Sharkey. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology* 14 (2012), 27–40.

[132] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. 2021. On the critical role of conventions in adaptive human-AI collaboration. *arXiv preprint arXiv:2104.02871* (2021).

[133] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th acm/ieee international conference on human-robot interaction (hri)*. IEEE, 219–226.

[134] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.

[135] Roland Siegwart, Illah Reza Nourbakhsh, and Davide Scaramuzza. 2011. *Introduction to autonomous mobile robots*. MIT press.

[136] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision*.

[137] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[138] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[139] JaYoung Sung, Henrik I Christensen, and Rebecca E Grinter. 2009. Robots in the wild: understanding long-term use. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 45–52.

[140] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems* 34 (2021), 251–266.

[141] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[142] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* (2023), 1–9.

[143] Sebastian Thrun. 2008. Simultaneous localization and mapping. *Robotics and cognitive approaches to spatial mapping* (2008), 13–41.

[144] Carme Torras. 2016. Service robots for citizens of the future. *European Review* 24, 1 (2016), 17–30.

[145] Katherine M Tsui and Holly A Yanco. 2013. Design challenges and guidelines for social interaction using mobile telepresence robots. *Reviews of Human Factors and Ergonomics* 9, 1 (2013), 227–301.

[146] Sherry Turkle. 2012. In Constant Digital Contact, We Feel'Alone Together'. *Alone Together* (2012).

[147] ubahnverleih. 2016. Nao Robot. https://commons.wikimedia.org/wiki/File:Nao_Robot_(Robocup_2016).jpg. Online; accessed 22-July-2023.

[148] Andries Van Dam. 1997. Post-WIMP user interfaces. *Commun. ACM* 40, 2 (1997), 63–67.

[149] Kazuyoshi Wada and Takanori Shibata. 2006. Robot therapy in a care house-its sociopsychological and physiological effects on the residents. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 3966–3971.

[150] Sameer Wagh, Divya Gupta, and Nishanth Chandran. 2019. SecureNN: 3-Party Secure Computation for Neural Network Training. *Proc. Priv. Enhancing Technol.* 2019, 3 (2019), 26–49.

[151] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. 2015. Holistic 3d scene understanding from a single geo-tagged image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3964–3972.

[152] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 2020. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In *Proceedings of (IROS) IEEE/RSJ International Conference on Intelligent Robots and Systems*. 10359 – 10366.

[153] Jacob O Wobbrock, Andrew D Wilson, and Yang Li. 2007. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. 159–168.

[154] Sarah Woods, Kerstin Dautenhahn, and Joerg Schulz. 2004. The design space of robots: Investigating children's views. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, 47–52.

[155] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. 2022. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12981–12990.

[156] Kuan Xu, Chen Wang, Chao Chen, Wei Wu, and Sebastian Scherer. 2022. Aircode: A robust object encoding method. *IEEE Robotics and Automation Letters* (2022).

[157] Holly A Yanco and Jill L Drury. 2002. A taxonomy for human-robot interaction. In *Proceedings of the AAAI fall symposium on human-robot interaction*. 111–119.

[158] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.

[159] Jian Yao, Sanja Fidler, and Raquel Urtasun. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 702–709.

[160] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5525–5534.

[161] Lai Sum Yim, Quang TN Vo, Ching-I Huang, Chi-Ruei Wang, Wren McQueary, Hsueh-Cheng Wang, Haikun Huang, and Lap-Fai Yu. 2022. WFH-VR: Teleoperating a Robot Arm to set a Dining Table across the Globe via Virtual Reality. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[162] Gareth Young, Hamish Milne, Daniel Griffiths, Elliot Padfield, Robert Blenkinsopp, and Orestis Georgiou. 2020. Designing mid-air haptic gesture controlled user interfaces for cars. *Proceedings of the ACM on Human-Computer Interaction* 4, EICS (2020), 1–23.

[163] Z22. 2014. iRobot Ava 500. https://commons.wikimedia.org/wiki/File:IRobot_Ava_500.jpg. Licensed under CC BY-SA 3.0 via Wikimedia Commons.

[164] Michelle Zhao, Reid Simmons, and Henny Admoni. 2022. The Role of Adaptation in Collective Human–AI Teaming. *Topics in Cognitive Science* (2022).

[165] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 1851–1858.

[166] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3357–3364.

A APPENDIX

ID	Robot Name	Link
1	Astro	https://www.aboutamazon.com/news/devices/meet-astro-a-home-robot-unlike-any-other
2	Atlas	https://www.bostondynamics.com/atlas
3	Ava	https://www.avarobotics.com/ava-webex
4	Baxter	https://robots.ieee.org/robots/baxter
5	Bluerov	https://bluerobotics.com
6	Da Vinci	https://www.intuitive.com/en-us/products-and-services/da-vinci/systems
7	DJI drone	https://www.dji.com/
9	Everyday robot	https://everydayrobots.com/
10	Fetch	https://fetchrobotics.com/freight-base-research/
11	Franka	https://www.franka.de/research
12	Kilobot	https://robotsguide.com/robots/kilobot/
13	LoCoBot	https://www.trossenrobotics.com/locobot-overview.aspx
14	Moley	https://www.moley.com/moley-kitchen
15	NAO	https://www.aldebaran.com/en/industries/government
16	Ohmni	https://ohmnilabs.com/products/ohmni-telepresence-robot/
17	PARO	http://www.parorobots.com/
18	Pepper	https://www.aldebaran.com/en/pepper
19	Phoenix	https://www.sanctuary.ai
20	Plato	https://cobiotx.unitedrobotics.group/en/plato
21	PR2	https://robots.ieee.org/robots/pr2/
22	Reachy	https://www.pollen-robotics.com
23	Roomba	https://www.irobot.com/en_US/why-irobot.html
24	Relay	https://www.relayrobotics.com/robots-in-action
25	Spot	https://www.bostondynamics.com/products/spot
26	Stretch	https://www.bostondynamics.com/products/stretch
27	TurtleBot	https://www.turtlebot.com
28	Unitree Go 1	https://shop.unitree.com/products/unitreeyushutechnologydog-artificial-intelligence-companion-bionic-companion-intelligent-robot-go1-quadruped-robot-dog
29	Vgo	http://www.vgocom.com

Table 2. A sample of robots and their weblinks.